

Communicated by Ellen Hildreth

---

## Interaction between Transparency and Structure from Motion

**Daniel Kersten**

*Department of Psychology, University of Minnesota,  
Minneapolis, MN 55455 USA*

**Heinrich H. Bülthoff**

**Bennett L. Schwartz**

**Kenneth J. Kurtz**

*Department of Cognitive and Linguistic Sciences, Brown University,  
Providence, RI 02912 USA*

It is well known that the human visual system can reconstruct depth from simple random-dot displays given binocular disparity or motion information. This fact has lent support to the notion that stereo and structure from motion systems rely on low-level primitives derived from image intensities. In contrast, the judgment of surface transparency is often considered to be a higher-level visual process that, in addition to pictorial cues, utilizes stereo and motion information to separate the transparent from the opaque parts. We describe a new illusion and present psychophysical results that question this sequential view by showing that depth from transparency and opacity can override the bias to see rigid motion. The brain's computation of transparency may involve a two-way interaction with the computation of structure from motion.

### 1 Introduction

---

One of the major challenges of vision research is to understand how the brain constructs a model of the visual environment from the pattern of changing retinal light intensities. With relatively few exceptions (Poggio *et al.* 1988; Barrow and Tenenbaum 1978), computational research has sought to first divide the problem into noninteracting modules such as surface color from radiance, shape from shading, or structure from motion (Land 1983; Horn and Brooks 1989; Ullman 1979; Martin and Aggarwal 1988). Consistent with the methodology of computer vision, current physiological and psychophysical research indicates modular and concurrent processing for some sources, such as motion, or form and color (Livingstone and Hubel 1987; Cavanagh 1987; Zeki 1978; Van Essen 1985).

In contrast to the modularity of vision research, it is phenomenally apparent that visual information is eventually integrated to provide a strikingly singular description of the visual environment. The simple unity of visual experience belies some difficult problems of neural computation. One problem is *cue integration*—the combination (possibly linear) of visual information from multiple sources, such as stereo and motion, to compute a single attribute such as depth. A second and theoretically more difficult problem is the *cooperative coupling* (typically nonlinear) of the perceptual representations of two or more scene attributes (such as surface depth and material property) to achieve the consistency required by the laws of image formation (Kersten 1991; Bühlhoff 1991; Bühlhoff and Yuille 1991). Because the outputs from two modules are not independent, algorithms require feedback between modules and thus are open to the problems of convergence and instability (Clark and Yuille 1990).<sup>1</sup>

We describe a new illusion, that has a bistable three-dimensional (3D) interpretation. The bistability is interpreted as the result of cooperative coupling between depth from motion and phenomenal surface transparency. Phenomenal transparency of a surface means we can see it and through it to another background surface. This illusion poses a problem for computational models of structure from motion because it seems to require cooperative coupling or strong fusion between representations of relative surface depth and surface material.

Motion provides information about relative depth relationships between surfaces in the world. Information for depth is available from motion parallax and motion disparity (Wallach and O'Connell 1953). Theoretical work has shown how structure from motion can be reconstructed with *a priori* structural biases, such as assuming that the object viewed is rigid (Martin and Aggarwal 1988).<sup>2</sup> Interactions between depth from motion and other depth sources, such as stereo and proximity luminance, have been studied before (Doshier *et al.* 1986; Nawrot and Blake 1989). With respect to transparency, it has recently been discovered that degree of transparency determines whether two superimposed and independently moving square wave grating patterns at right angles to each other are seen as moving in a single direction or in two independent directions (Ramachandran 1989; Stoner *et al.* 1990). In these experiments, when the luminance of the intersection of the two gratings was consistent with that derived from a physically transparent grating, the motion of the two gratings was seen to be independent—they appeared to be

<sup>1</sup>Our distinction between integration and cooperative coupling closely corresponds to the distinction between weak data fusion and strong fusion with recurrency made by Clark and Yuille (1990).

<sup>2</sup>Motion parallax typically refers to the differences in image speed, due to viewpoint changes, between points at different distances under perspective projection. Because the same retinal optic flow pattern can be induced by moving the object, the term motion parallax is sometimes used in this case too. Structure from motion typically refers to the reconstruction of object geometry due to its motion under orthographic or perspective projection.

sliding over each other. However, if the intersection luminance was not consistent with a physically plausible transparency or with occlusion, the gratings appeared to cohere as a single pattern moving in a unique direction. The conclusion was that motion detecting mechanisms may have tacit knowledge of the physics of transparency. In these studies, the relationship between depth and transparency is only implicit. Although the transparency of a surface implies that it is closer than the surface it covers, one would like to know whether transparency can provide specific depth information that could affect three-dimensional structure from motion. It turns out that particular intensity relationships not only determine whether transparency is seen (Metelli 1974; Richards and Witkin 1979; Beck *et al.* 1984), but as is shown below, also bias which of two overlapping surfaces is seen in front. We call this depth from transparency. How do depth from motion and transparency interact? In particular, when depth from motion and depth from transparency contradict, which takes precedence—motion or transparency information?

## 2 Perceptual Observations

---

In an attempt to answer the above questions we simulated an object consisting of two square planar parallel surfaces that could rigidly rock back and forth at  $\pm 40^\circ$  about a common vertical axis midway between them (see Fig. 1 for more details). The planes could be seen as square when in a head-on view, but typically appeared trapezoidal due to perspective. Either the top or bottom face could be made to appear in front of the other depending on apparent transparency and depth from motion. The particular intensity relationships of the four regions bias the apparent transparency of a face, and thus determine the relative depth of the front and back planes. The motion parallax, together with a bias toward rigidity (Wallach *et al.* 1953) also affects the depth one sees.

In the following we will describe the basic perceptual phenomena, and then detail the results of some quantitative psychophysical measurements. In all three of the demonstrations described below, the rigid motion is described as being consistent with the bottom face being in front of the top face and only the intensities of the various regions are changed. As described in Section 3, the basic observations are unaffected by whether the top or bottom face is in front.

First we looked at the case in which both surfaces have zero transparency—that is, they are both opaque with the bottom square in front, and partially occluding the top (Fig. 2a). When the object was rocked back and forth, not surprisingly, observers saw rigid motion that was consistent with both the motion and occlusion cues. The surfaces do and appear to share a common rotation axis that is behind the bottom square, and in front of the top square. Next the intensity of the center patch of overlap was changed to match the top face. Thus the top patch appeared to

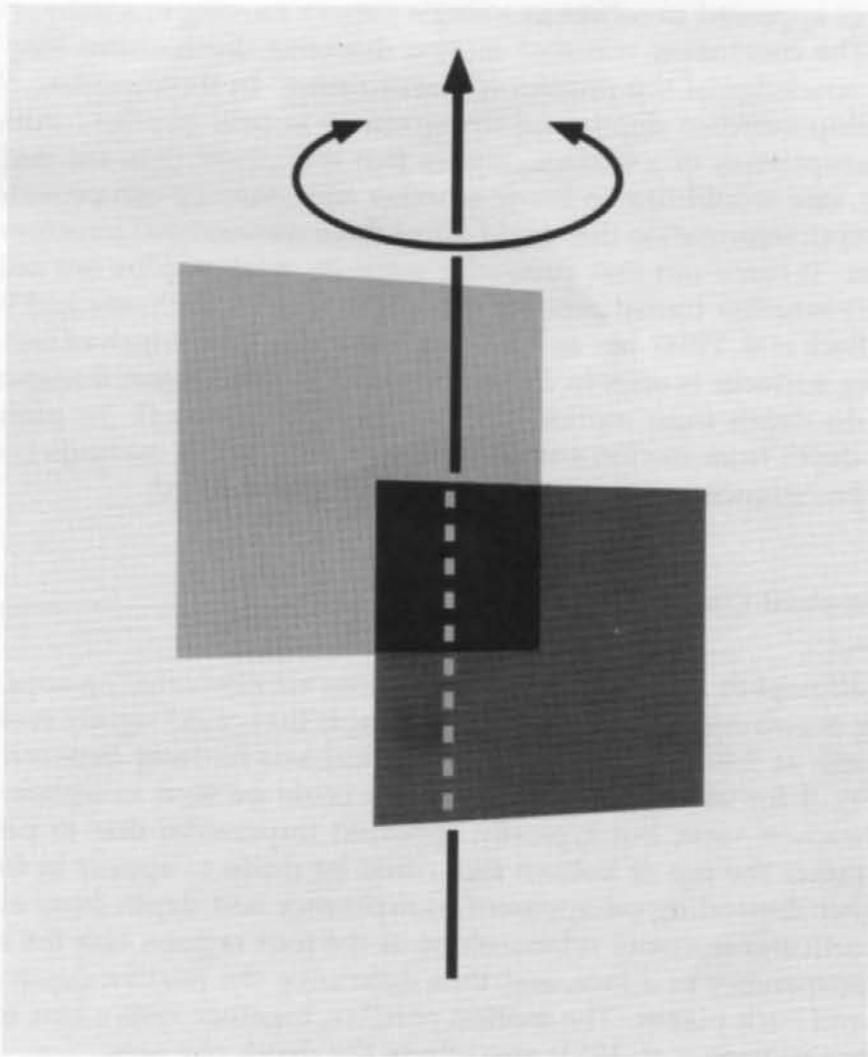


Figure 1: Animated sequences of images corresponding to a perspective view of two rigidly coupled planar faces (each a simulated  $5 \times 5$  cm square) were generated with a Macintosh II computer and displayed on a CRT monitor with a 256 gray-level capacity. The object was rocked back and forth rigidly about the vertical axis passing between the two surfaces and through a point equidistant to both. Like the Necker cube, which is an orthographic projection of a wire cube, a particular image frame can give rise to an ambiguous depth percept: the top face can appear in front or behind the bottom face. The planes oscillated sinusoidally over an amplitude of  $\pm 40^\circ$  at 0.48 Hz. The front and back faces were separated by a simulated depth of 6 cm. The centers of the two faces were separated by 2.5 cm in the horizontal and vertical directions. The distance between the point equidistant between the two faces and the observer's eye-point was 57 cm. There were 21 frames per period.

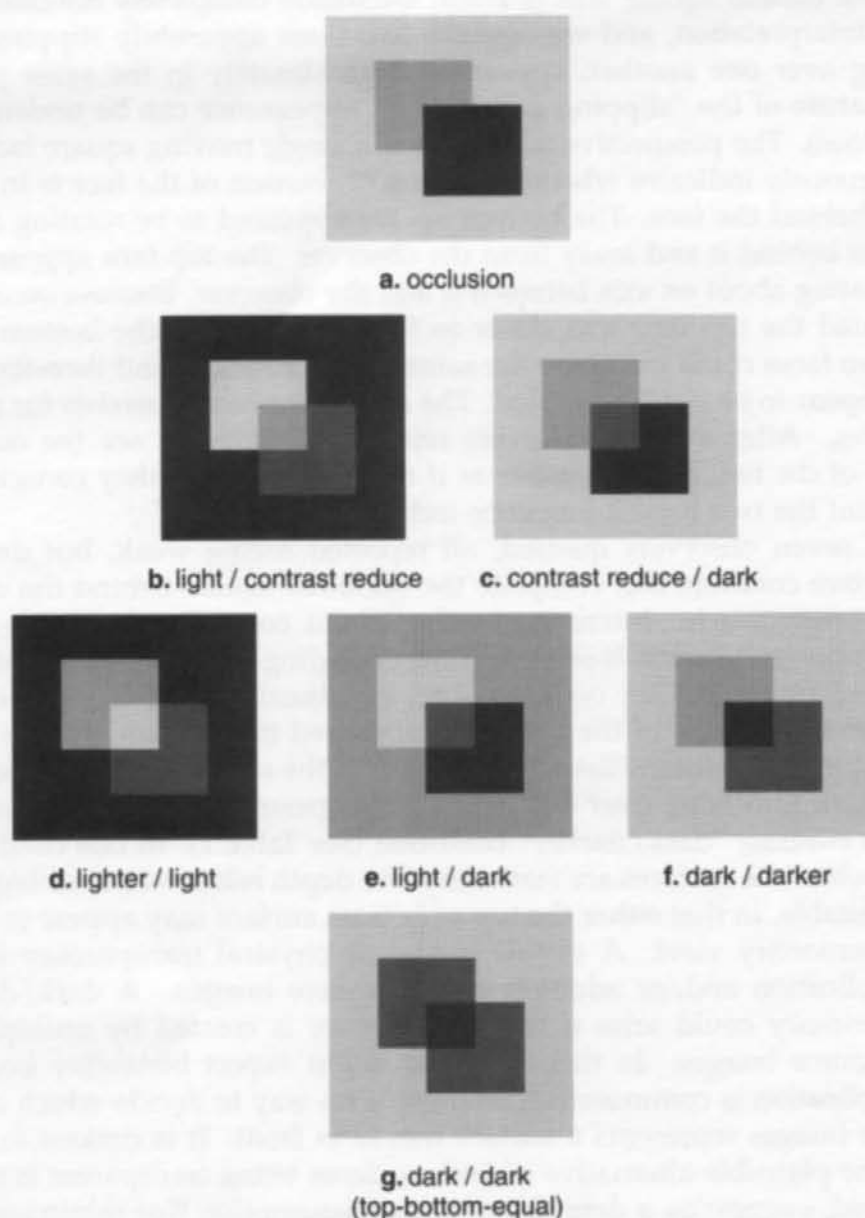


Figure 2: Different transparency types. The first and second words on the label for a transparency type indicate how the top and bottom faces affect the brightness of the patches they cover, respectively. For example, the dark/darker transparency means that both the top and bottom faces darkened what they cover, and that the bottom one was darker than the top. Note that in the actual animation, the nearer square appeared slightly larger because of the perspective projection. Note how the contrast reducing faces (bottom square in b and top square in c) tend to appear nearer the observer.

occlude the bottom, in contradiction to the rigid motion, which indicated that the bottom square was in front. Occlusion completely inhibited the rigid interpretation, and we saw the two faces apparently slipping and sliding over one another, appearing approximately in the same plane. The nature of the "slipping and sliding" appearance can be understood as follows. The perspective projection of a single moving square face unambiguously indicates whether the axis of rotation of the face is in front of or behind the face. The bottom square appeared to be rotating about an axis behind it and away from the observer. The top face appeared to be rotating about an axis between it and the observer. Because occlusion indicated the top face was closer to the observer than the bottom face, the two faces could not share the same axis of rotation, and therefore did not appear to be rigidly coupled. The nonrigid percept persists for many minutes. After awhile, observers report that they can see the outside edges of the two surfaces move as if rigidly coupled if they consciously discount the two local T-junctions indicating occlusion.<sup>3</sup>

Of seven observers queried, all reported seeing weak, but definite subjective contours that complete the occluded square behind the center overlapping patch. Interestingly, these faint contours are visible even when nonrigid motion is seen, as if the occluding patch were transparent.

Next we relaxed the occlusion cue, by adjusting the intensities of the patches so that one of the two faces appeared transparent. In one case, we adjusted the intensities so that either of the surfaces could appear to be a dark film lying over a light gray background, referred to below as a high contrast "dark/darker" condition (see Table 1). In this condition, even when the surfaces are stationary, the depth relations are ambiguous and bistable, in that either the top or bottom surface may appear in front in a stationary view. A simple model of physical transparency is the multiplication and/or addition of two source images. A dark/darker transparency could arise if the transparency is created by multiplying two source images. In this case, one might expect bistability because multiplication is commutative, so there is no way to decide which of the source images represents a surface that is in front. It is curious to note that the plausible alternative of both surfaces being transparent is never reported, suggesting a default perceptual assumption that minimizes the number of transparent surfaces. One can also adjust the intensities of the top and bottom squares to be equal. In this case the only biases to favor seeing a plane in front are to prefer the bottom over the top, and the larger over the smaller (Fig. 2g). In either the dark/darker or the top-bottom-equal condition, when the two faces were rocked back and forth, we saw a striking bistability. With the bottom face in front, we saw both planes rigidly rocking back and forth with the bottom face appearing transparent, and the top face opaque. After watching this for 2 to 30 sec, suddenly the top face would appear in front and then the

<sup>3</sup>A T-junction occurs where the edge of the occluder covers the edge of the occluded surface. An X-junction is the image point where transparent and opaque contours cross.

Table 1: Intensity Values (cd/m<sup>2</sup>) for the transparency types.<sup>a</sup>

Transparency type	Top patch	Center patch	Bottom patch	Background Background	Contrast (%)
Dark/dark with top-bottom-equal	26	16	26	51	-24
Occlusion	38	16	16	51	0
Dark/darker (HC)	38	16	26	51	-24
Contrast reduce/dark (HC)	38	26	16	51	24
Light/dark (HC)	51	26	16	38	24
Light/contrast reduce (HC)	51	38	26	16	19
Lighter/light (HC)	38	51	26	16	32
Dark/darker (LC)	38	16	19	51	-8.6
Contrast reduce/dark (LC)	38	26	23	51	6.1
Light/dark (LC)	51	26	23	38	6.1
Light/contrast reduce (LC)	51	38	34	16	5.6
Lighter/light (LC)	38	51	46	16	5.2

<sup>a</sup>In addition to occlusion and a dark/dark top-bottom-equal transparency, the choice of transparency types was motivated by a consideration of the possible transparencies one can generate permuting four intensities. There are twenty-four possible permutations, but these can be reduced to just six by excluding top/bottom symmetry and the physically implausible contrast reversing and contrast enhancing pairs. Of these six, two involve faces that both darkened the underlying surfaces, so one was eliminated, leaving five. To further increase the range of transparency types, we also added five stimuli in which the Michelson contrast of the lower right-hand corner of the central patch was smaller. The high (HC) and low contrast (LC) groups had contrasts whose absolute values were above 19% and below 8.6%, respectively.

perceived motion was one of two faces slipping and sliding over each other. Simultaneous with this reversal of depth, there was an exchange of surface property—the top face now appeared transparent and the bottom opaque. In the top-bottom-equal condition, there is no depth from transparency bias to see one or the other face in front. Nevertheless, at a given moment, the visual system makes a commitment to one of two plausible relative depth and transparency assignments. The specific default assignment of depth, which lasts for a while and then changes, is similar to what happens when viewing a stationary Necker cube. Our demonstration, in addition, clearly demonstrates that relative depth interacts with apparent surface transparency.

In a third demonstration, we sought a condition intermediate between the symmetric transparency of a dark/dark combination and complete occlusion by constructing a transparent overlay that appears diaphanous. A diaphanous transparent square has both additive and multiplicative components that bias its relative depth to be in front of the other square. This can be physically realized by a finely perforated screen whose holes are below the spatial resolution limit and that transmits a fraction of the light coming from behind, and reflects a fraction coming from the front (Kersten 1991; Richards and Witkin 1979). Consistent with the interpretation of a perforated screen, a film that reduces the contrast of the edges

it overlays by lightening the darker region, and darkening the lighter, without changing contrast polarity tends to be seen in front (Fig. 2b and c). In the demonstration, the top square was made to appear contrast reducing. The bottom square was made to appear as a dark patch behind the contrast-reducing top square (the high contrast "contrast reduce/dark" condition in Table 1, Fig. 2c). When the two faces were rocked back and forth, we saw the wrong motion. Just as in the case of occlusion, the surfaces appeared to slip nonrigidly over one another with the top face appearing in front. After several seconds of observation, suddenly rigid motion is seen at which time the top contrast-reducing square is seen behind a dark bottom film. Again there was a simultaneous and unambiguous reversal of apparent transparency—the contrast-reducing top square suddenly appeared opaque and behind a dark film at the bottom.

Our paradigm is potentially useful for studying other interactions between structure from motion and transparency. For example, informal observations have shown that connecting the corners of the two faces with lines (e.g., a wire-frame outline of the cube) can override the non-rigid interpretation induced by weak perspective information. Further, even without connecting wires, an orthographic projection eliminates the nonrigid interpretation. Under orthographic projection, the appearance of the two faces flips between two rigid interpretations, like a Necker cube. In this case, the transparency biases which of the two rigid interpretations are seen.

### 3 Psychophysics: Response Time vs. Face-in-Front Bias \_\_\_\_\_

To quantify the interaction between transparency cues on depth and structure from motion, we made measurements of the reaction time to see rigid motion conditional on the perceived depth relations seen in an initial static view. The time to see rigid motion was measured in two basic conditions in which the initial depth perception, based on transparency, could either conflict (*inconsistent* condition) or agree (*consistent* condition) with the subsequent 3D motion. The experimental set-up was as before.

By specifying the gray levels of the four image regions, it was possible to vary the transparency, and thus produce a range of biases for whether a face of a particular transparency type appeared in front. This range of biases can be seen in the abscissa of Figure 3 where the proportion of times a particular face (e.g., the darker of a dark/darker combination) is seen in front (the face-in-front bias) ranges from 50 to 100% of the time depending on the transparency type for a given observer. We chose 12 different transparency types summarized in Table 1. The notation for the transparency type indicates how the top and bottom patches affect the brightness of the background. On half of the trials, the top face

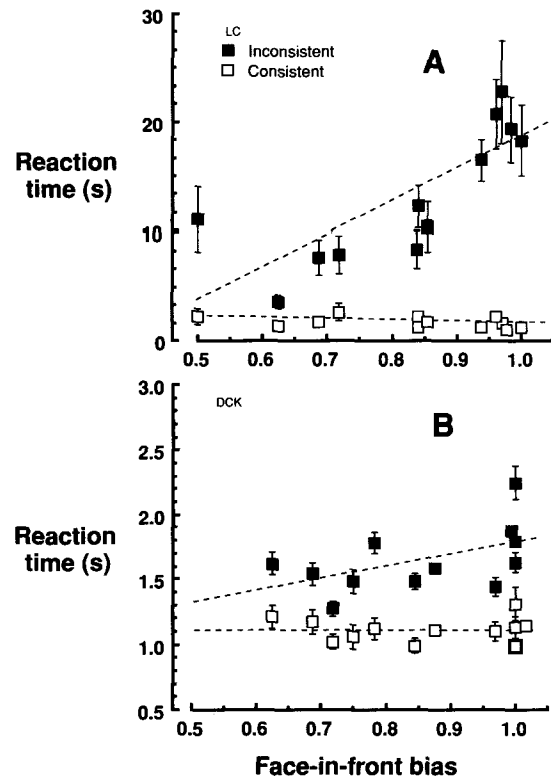


Figure 3: Mean time ( $\pm$ SEM) to see rigid motion plotted against the face-in-front bias for two observers (A and B). The face-in-front bias is the proportion of times a particular face appeared in front in the initial static view (e.g., the contrast-reducing face tended to appear in front of whatever face it overlapped). Results from the 12 different transparency conditions are shown. Each point is the mean of 16 measurements, averaged over conditions in which the top and bottom intensities were exchanged.

was in front of the bottom face (front-top), as defined by the subsequent motion, and on the other half of the trials, it was behind the bottom face (front-bottom). Because the perspective view made the image of the front patch larger than the back, we balanced for a possible bias by showing the observers the stimuli with the top and bottom intensities "normal" or "exchanged" for each of the front-top and front-bottom conditions. Because we could not guarantee that a given transparency condition would













